City University of New York (CUNY)
CUNY Academic Works

Open Educational Resources Queensborough Community College

2020

Clear-Sighted Statistics: Module 11: Confidence Intervals (slides)

Edward Volchok
CUNY Queensborough Community College

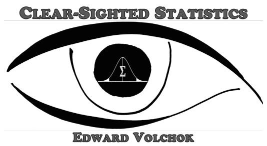
How does access to this work benefit you? Let us know!

More information about this work at: https://academicworks.cuny.edu/qb_oers/154
Discover additional works at: https://academicworks.cuny.edu

This work is made publicly available by the City University of New York (CUNY).
Contact: AcademicWorks@cuny.edu

Confidence Intervals

Module 11



cc by nc

0

Arthur C. Nielsen, Jr.
(1919 - 2011)

"Too many business people assign equal validity to all numbers printed on paper. They accept numbers as representing the Truth and find it difficult to work with the concept of probability. They do not see a number as a shorthand for range that describes our actual knowledge of the underlying condition."

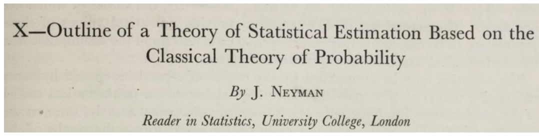
-- Arthur C. Nielsen, Jr.

Arthur C. Nielsen, Jr. "Statistics in Marketing." In George Easton, Harry V. Roberts, and George, George C. (Eds.): *Making Statistics More Effective in Schools of Business*. Chicago, IL: University of Chicago Graduate Schools of Business, 1986.

cc by nc

1

In the 1930s, Jerzy Neyman introduces Confidence Intervals


X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. NEYMAN

Reader in Statistics, University College, London

Jerzy Neyman
1894 - 1981

cc by nc

2

Confidence intervals are used to estimate unknown parameters using sample statistics

cc by nc

3

Lecture Objectives

- Define confidence intervals, point estimates, levels of confidence, upper and lower confidence limits, and the margin of error (MoE)
- Construct confidence intervals for µ when σ is known using z-values
- Construct confidence intervals for µ when σ is unknown using t-values
- Construct confidence intervals for a population proportion, π, using z-values
- Apply the Finite Population Correction (FPC) factor (when needed)

4

Confidence Interval Questions

- How much debt has the average college student incurred?
- How many high school and middle school students vape?
- What proportion of the American public supports stricter gun control or fewer restrictions on a woman's right to an abortion?
- What proportion of automobile buyers intend to purchase an electric or hybrid car in the next four years?
- What proportion of people who have health insurance from their employer are satisfied with their insurance plan?

5

Confidence Interval Terms

Key Terms

6

Confidence Interval (CI)

- A confidence interval or interval estimate is a range of values obtained from a sample that is likely to contain the parameter we seek to estimate
- Confidence intervals contain the actual parameter in the long run a certain percentage of time
- We will construct two-sided or two-tailed intervals

7

Point Estimate

Sample statistic, $\bar{X}$ or p, upon which the confidence interval is based

$\bar{X}$ = sample mean

p = sample proportion

The confidence intervals are split evenly around the point estimate, half the interval will be above the point estimate and half will be below

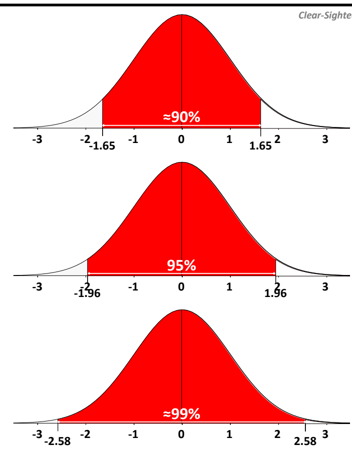
Confidence Level (CL)

States the degree of certainty that the estimate of the unknown parameter is included in the confidence interval

95% level of confidence used most frequently

Occasionally 99% or 90% confidence intervals are used

The higher the confidence level, the wider the CI



Confidence Levels and Significance Levels, α

Confidence levels and significance levels, or α, are closely related

1 – CL = α, 1 – α = CL

Significance Levels are used in NHST

Excel uses Significance Levels when calculating CI

CL and α

	Confidence Level (CL)	Significance Level (α)	Found By
95%	95%	5%	1 - CL
99%	99%	1%	1 - CL
90%	90%	10%	1 - CL

12

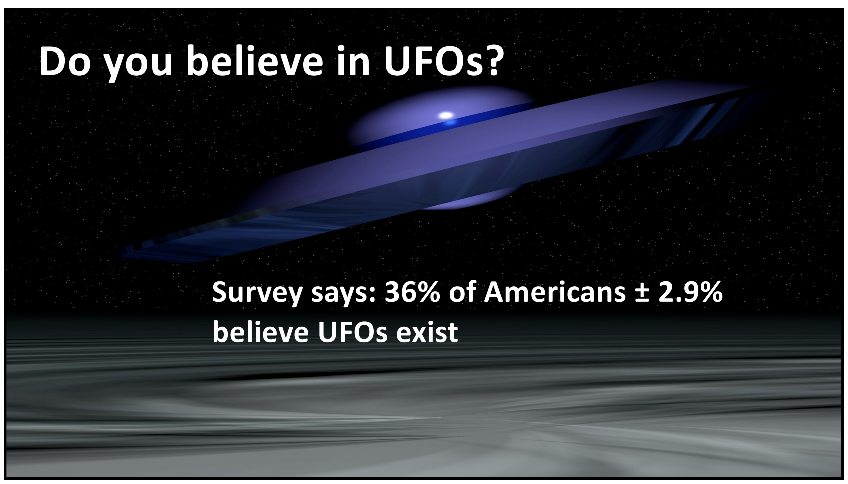
Margin of Error (MoE)

Margin of error is the width of the confidence interval

The distance the confidence interval extends above and below the point estimate

Should be shown when survey results are presented

13



14

National Geographic's Survey on UFOs

Survey conducted May 21-29, 2012

Nationwide random sample of 1,114 Americans

36% is the point estimate

2.9% is the Margin of Error

15

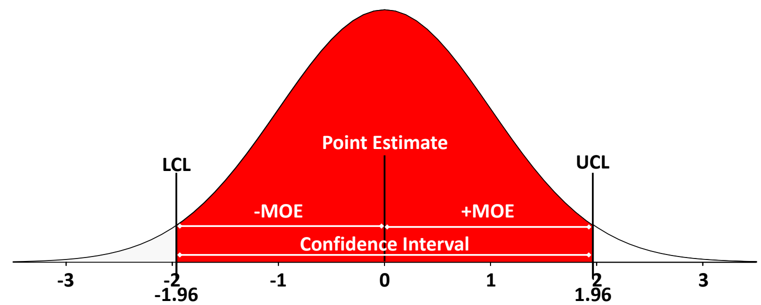
Confidence Limits (LCL & UCL)

Confidence limits are the upper and lower limits of the confidence interval

Abbreviated as UCL and LCL

Confidence Limit	Found By
Upper Confidence Limit (UCL)	Point Estimate + MoE
Lower Confidence Limit (LCL)	Point Estimate - MoE

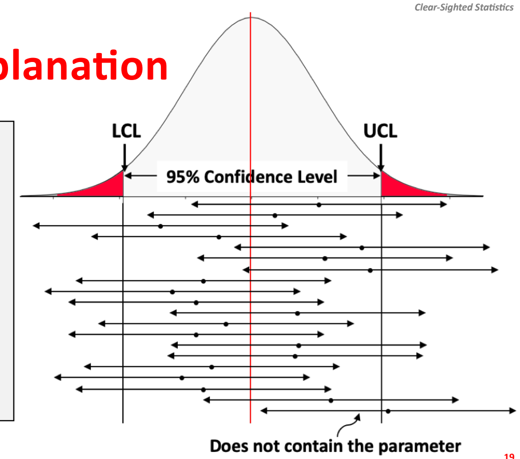
Point Estimate & MoE



Confidence Interval Basics

A common explanation

95% CI contains the estimated parameter 95% of the time when the surveys are repeated



Why this is a misconception

This would only be true if the initial estimate is exactly equal to the parameter

The average probability that a first 95% confidence interval capturing the statistic from the next sample is only around 83.4%*

*Geoff Cumming, Jennifer Williams, and Fiona Fidler. "Replication and Researchers' Understanding of Confidence Intervals and Standard Error Bars." *Understanding Statistics*. Vol. 3, October 2004. pp. 299-311.

20

Three Factors Affect the MoE
(width of the CL)

Confidence Level (CL)	Sample Size (n)	Data Variability
• The higher the CL, the wider the CI	• Bigger the sample size, the narrower the CI	• The more variable the data, the wider the CI

21

CI for μ when σ is known

22

Confidence Interval Formula

$$CI = \bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

Where: $\bar{X}$ = the sample mean, is the point estimate for the population mean, μ
σ = the population standard deviation
z = the z-value for the selected confidence level
n = the number of observations in the sample
σ/√n is the standard error of the mean

23

z-Values to use

Excel provides greater precision than a critical values table

CL	CV Table	Value Used by Excel
90%	1.65	1.645
95%	1.96	1.960
99%	2.58	2.576

Excel uses values worked out to 15 decimal places

Example

What is the monthly rent for a 1 Bedroom apartment in Manhattan?

Without a census, we cannot know μ with certainty

We can estimate the population mean, μ, using sample data and CI

Sample Statistics

$$\$2{,}520.22 \pm 1.96\frac{\$884.92}{\sqrt{100}} =$$

$$\$2{,}520.22 \pm 1.96\frac{\$884.92}{10} =$$

$$\$2{,}520.22 \pm 1.96\left(88.94\right) =$$

$$\boxed{\$2{,}520.22 \pm \$174.44}$$

$\bar{X}$ = \$2,520.22
σ = \$889.38
n = 100
CL = 95%

$$CI = \bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

Excel's CONFIDENCE.NORM

Slightly more precise

	A	B	C	D	E
1	One Bedroom	Monthly Rent	Confidence Level	0.95	Formula
2		\$2,400	Sample Mean	\$2,520.22	=AVERAGE(B2:B101)
3		\$2,699	Pop. Std. Dev.	\$884.92	=STDEV.P(B2:B101)
4		\$1,399	n	100	=COUNT(B2:B101)
5		\$1,499	z-value	1.96	=ABS(NORM.S.INV(0.025))
6		\$2,380	Std. Error	88.49	=D3/SQRT(D4)
7		\$1,800	MoE	\$173.44	=D5*D6
8		\$1,550	LCL	\$2,346.78	=D2-D7
9		\$2,900	UCL	\$2,693.66	=D2+D7
10		\$2,600	Alpha	0.05	=1-D1
11		\$3,000	Excel	\$173.44	=CONFIDENCE.NORM(D10,D3,D4)
12		\$2,600			

±\$173.44 vs. ±\$174.44

Problem with this CI

If we do not know μ, it is unlikely that we know σ

We cannot use z-values when σ is unknown

We need a new distribution!

28

William Gosset's Student-t

29

Normal Distribution Assumptions

- Shape of the population may be unknown, but the number of observations is ≥ 30
- If n is ≥ 30, Central Limit Theorem states that the sampling distribution is normal
- σ is known

30

Meet William Gosset

- Chemist at Guinness Brewery
- Worked with tiny samples of grains
- He could not invoke CLT and use z-values
- Developed the student-t distribution for use when the Normal Distribution assumptions are not met
- Today student-t is used more often and z-values

William Gossett
(1876 – 1937)

31

R. A. Fisher
1890 - 1962

In 1912, Ronald A. Fisher set the mathematical foundation for the t-distribution

32

Student-t Distribution

- Continuous and symmetrical like the normal distribution
- Flatter peaks
- Thicker tails

33

Student-t is defined by the…

- Sample Mean, $\bar{X}$
- Sample Standard Deviation, s
- Degrees of Freedom, (n – the number of independent samples)
- Student-t is used for 1 or 2 samples

34

Student-t Table

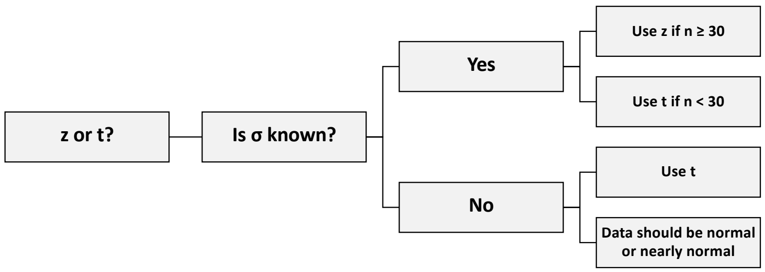
- Critical values for t can be found using a CV table, or
- Microsoft Excel

	Confidence Level (CL)					
	80%	90%	95%	98%	99%	99.9%
	α — One-Tailed Test					
	0.10	0.05	0.025	0.01	0.005	0.0005
	α — Two-Tailed Test					
df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
31	1.309	1.696	2.040	2.453	2.744	3.633
32	1.309	1.694	2.037	2.449	2.738	3.622
33	1.308	1.692	2.035	2.445	2.733	3.611
34	1.307	1.691	2.032	2.441	2.728	3.601
35	1.306	1.690	2.030	2.438	2.724	3.591

	Confidence Level (CL)					
	80%	90%	95%	98%	99%	99.9%
	α — One-Tailed Test					
	0.10	0.05	0.025	0.01	0.005	0.0005
	α — Two-Tailed Test					
df	0.20	0.10	0.05	0.02	0.01	0.001
36	1.306	1.688	2.028	2.434	2.719	3.582
37	1.305	1.687	2.026	2.431	2.715	3.574
38	1.304	1.686	2.024	2.429	2.712	3.566
39	1.304	1.685	2.023	2.426	2.708	3.558
40	1.303	1.684	2.021	2.423	2.704	3.551
41	1.303	1.683	2.020	2.421	2.701	3.544
42	1.302	1.682	2.018	2.418	2.698	3.538
43	1.302	1.681	2.017	2.416	2.695	3.532
44	1.301	1.680	2.015	2.414	2.692	3.526
45	1.301	1.679	2.014	2.412	2.690	3.520
46	1.300	1.679	2.013	2.410	2.687	3.515
47	1.300	1.678	2.012	2.408	2.685	3.510
48	1.299	1.677	2.011	2.407	2.682	3.505
49	1.299	1.677	2.010	2.405	2.680	3.500
50	1.299	1.676	2.009	2.403	2.678	3.496
51	1.298	1.675	2.008	2.402	2.676	3.492
52	1.298	1.675	2.007	2.400	2.674	3.488
53	1.298	1.674	2.006	2.399	2.672	3.484
54	1.297	1.674	2.005	2.397	2.670	3.480
55	1.297	1.673	2.004	2.396	2.668	3.476
56	1.297	1.673	2.003	2.395	2.667	3.473
57	1.297	1.672	2.002	2.394	2.665	3.470
58	1.296	1.672	2.002	2.392	2.663	3.466
59	1.296	1.671	2.001	2.391	2.662	3.463
60	1.296	1.671	2.000	2.390	2.660	3.460
61	1.296	1.670	2.000	2.389	2.659	3.457
62	1.295	1.670	1.999	2.388	2.657	3.454
63	1.295	1.669	1.998	2.387	2.656	3.452
64	1.295	1.669	1.998	2.386	2.655	3.449
65	1.295	1.669	1.997	2.385	2.654	3.447
66	1.295	1.668	1.997	2.384	2.652	3.444
67	1.294	1.668	1.996	2.383	2.651	3.442
68	1.294	1.668	1.995	2.382	2.650	3.439
69	1.294	1.667	1.995	2.382	2.649	3.437
70	1.294	1.667	1.994	2.381	2.648	3.435

35

Deciding on using z or t



CC BY-NC-SA

CI for μ when σ is unknown

Student t Distributions

CC BY-NC-SA

Confidence Interval Using t

$$CI = \bar{X} \pm t\frac{s}{\sqrt{n}}$$

Where: $\bar{X}$ = the sample mean, is the point estimate for the population mean, μ
s = the sample standard deviation
t = the t-value for the selected confidence level
n = the number of observations in the sample
σ/√n is the standard error of the mean

CC BY-NC-SA

Sample Statistics

Finding the Critical Value for t using Excel

	A	B	C
1	n	100	
2	df	99	=B1-1
3	CL	0.95	
4	α	0.05	=1-B3
5	CV	1.984	=TINV(B4,B2)

$$CI = \bar{X} \pm t\frac{s}{\sqrt{n}}$$

CC BY-NC-SA

Sample Statistics

$$\$2{,}520.22 \pm 1.984\frac{\$889.39}{\sqrt{100}} =$$

$$\$2{,}520.22 \pm 1.984\frac{\$889.39}{10} =$$

$$\$2{,}520.22 \pm 1.984(88.94) =$$

$$\boxed{\$2{,}520.22 - \$176.45}$$

$\bar{X}$ = $2,520.22
s =$889.38
n = 100
df = 99
CL = 95%
t = 1.984

$$CI = \bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

Excel's CONFIDENCE.T Function

12		$2,600	C	D	E
13		$2,400	Using t-values		
14		$1,750	Sample Mean	$2,520.22	=AVERAGE(B2:B101)
15		$1,675	Sample Std. Dev.	$889.38	=STDEV.P(B2:B101)
16		$1,899	n	100	=COUNT(B2:B101)
17		$1,595	t-value	1.984	=ABS(T.INV(0.025,D4-1))
18		$2,000	Std. Error	88.94	=D15/SQRT(D16)
19		$3,200	MoE	$176.47	=D17*D18
20		$4,350	LCL	$2,343.75	=D14-D19
21		$2,000	UCL	$2,696.69	=D14+D19
22		$1,699	Alpha	0.05	=1-D1
23		$1,675	Excel	$176.47	=CONFIDENCE.T(D22,D15,D16)

Using t gives a wider CI than z

t:
$2,520.22 ± $176.47

z:
$2,520.22 ± $174.44

CI for the population proportion, π

What are proportions?

Proportions are represented as a fraction, decimal, ratio, or percentage of the part of a population or sample that has a certain characteristic

Proportions are binary (success or failure)

CC BY NC SA

Formula for Proportions

$$p = \frac{X}{n}$$

Where: X = the random variable
n = the number of observations in the sample
p = the sample proportion

CC BY NC SA

Typical proportion questions

The proportion of Americans who think President Trump is a racist

The proportion of men with male pattern baldness

The proportion of Americans who follow the paleo diet

CC BY NC SA

CI for the Proportion

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

Note: We do not use t-values for CI for proportions

CC BY NC SA

Standard Error of the Proportion (SEP)

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

48

Gallup Poll: Legalization of Marijuana

October 2019: 66% of Americans favor the legalization of marijuana for recreational use

Party	Number Surveyed	Favor Legalization	%
Republican	393	199	50.64%
Independent	655	444	67.79%
Democrat	450	352	78.22%
Total	1,498	995	66.42%

49

Sample Proportion for Republicans

$$p = \frac{X}{n} = \frac{199}{393} = 0.5064$$

50

95% CI for Republicans

$$0.5064 \pm 1.96\sqrt{\frac{0.5064(1-0.5064)}{393}} =$$

$$0.5064 \pm 1.96\sqrt{\frac{0.24996}{393}} =$$

$$0.5064 \pm 1.96(0.02522) =$$

$$\boxed{0.5064 \pm 0.0494}$$

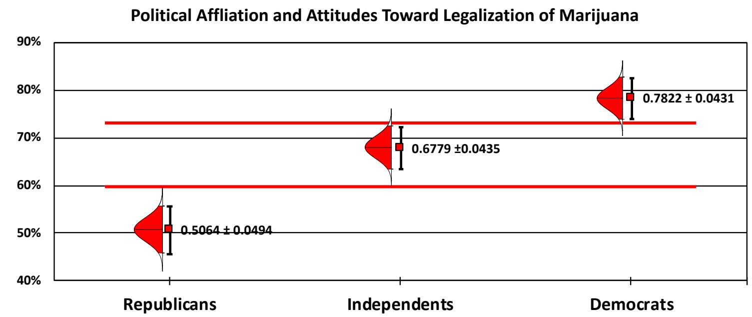
$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

51

How do Republicans compare to Democrats and Independents?

	A	B	C	D	E	F	G	H	I	J	K	L
1	Party	X	n	p	z	(1 - p)	p(1 - p)	p(1 - p)/n	Std. Err.	MoE	LCL	UCL
2	Republicans	199	393	0.5064	1.96	0.494	0.24996	0.00064	0.02522	0.0494	0.4569	0.5558
3	Independents	444	655	0.6779	1.96	0.322	0.21836	0.00049	0.02218	0.0435	0.6344	0.7213
4	Democrats	352	450	0.7822	1.96	0.218	0.17035	0.00048	0.02200	0.0431	0.7391	0.8253
5												
6	Gallup Poll											
7	23-Oct-19											
8	https://news.gallup.com/poll/267698/support-legal-marijuana-steady-past-year.aspx											

Charting the Confidence Intervals



Finite Population Correction Factor

(FPC Factor)

Finite Population Correction Factor (FPC)

- Used for small populations
- Makes CI more precise
- Used when samples are ≥ 5% of the population
- Sampling performed without replacement

FPC Formula

$$FCF = \sqrt{\frac{N - n}{N - 1}}$$

Where: n = the number of observations in the sample
N = the number of observations in the proportion

CI Formula With FPC

$$\bar{X} \pm t \frac{s}{\sqrt{n}} \left(\sqrt{\frac{N - n}{N - 1}} \right)$$

FPC Example

Lara is secretary of her high school graduating class

She is preparing the 20th reunion

She wants to estimate the average number of children her classmates have

She takes a sample

CL = 95%
N = 150
n = 25
$\bar{X}$ = 1.60
s = 1.20
df = 24
t = 2.064

$$\bar{X} \pm t \frac{s}{\sqrt{n}} \left(\sqrt{\frac{N - n}{N - 1}} \right)$$

The Calculation

$$1.60 \pm 2.064 \frac{1.20}{\sqrt{25}} \left(\sqrt{\frac{150 - 25}{150 - 1}} \right) =$$

$$1.60 \pm 2.064 \frac{1.20}{5} (0.916) =$$

$$1.60 \pm 2.064 (0.24)(0.916) =$$

$$1.60 \pm 0.454$$

$$LCL = 1.146$$

$$UCL = 2.054$$

$$\bar{X} \pm t \frac{s}{\sqrt{n}} \left(\sqrt{\frac{N - n}{N - 1}} \right)$$

60

Except where otherwise noted *Clear-Sighted Statistics* is licensed under
a Creative Commons License. You are free to share derivatives of this work
for non-commercial purposes only. Please attribute this work to Edward Volchok.

CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK

61